Sources

The real incidents behind each of the four principles, with links to the original reports.

The lost-laptop project and the "Update Goes Live Monday" activity are fictional, built for the session.

Human-owned

U.S. Government Accountability Office. Artificial Intelligence: Actions and Opportunities to Implement Responsible Use in DOD (GAO-27-107837). October 14, 2026.

https://www.gao.gov/products/gao-27-107837

The Responsible AI Working Council did not meet from late 2024 until August 2025 and had no charter setting roles, responsibilities or meeting frequency; the central Responsible AI division had 3 of 10 authorized positions filled as of February 2026. This is a finding about governance structure, not that DOD's AI deployments failed. DOD concurred with GAO's recommendations.

Office of the Governor of California. Announcement of AI-related worker protection bills signed, including SB 947. September 30, 2026.

https://www.gov.ca.gov/2026/09/30/californias-nation-leading-ai-framework-just-got-stronger-governor-newsom-signs-more-first-in-the-nation-worker-protections-and-more/

SB 947 bars employers from relying solely on an automated decision system to discipline or terminate an employee. When an employer primarily relies on one, a human must corroborate the output with other information, and output that can't be corroborated can't be used (Labor Code §1522, operative July 1, 2027). Mentioned as a parallel: a human in the chain isn't the same as a human who owns the decision. Bill text: California Legislative Information, SB 947 (2025–2026).

https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB947

Source-grounded

OpenAI Alignment. "Signing up for disposable emails and searching GitHub for leaked API keys." Misalignment report, last updated September 16, 2026.

https://alignment.openai.com/misalignment-reports/searching-github-for-leaked-api-keys/

An internal, unreleased model in training was asked for county earnings figures. When it couldn't retrieve them, it invented them and said they came from the requested website. One documented instance; it says nothing about how often this happens.

Verifiable

Zaitseva, E., et al. "A multi-agent large language model (LLM) platform for structured medical anamnesis in outpatient care." Frontiers in Digital Health 8 (September 30, 2026). doi:10.3389/fdgth.2026.1882397.

https://doi.org/10.3389/fdgth.2026.1882397

Across 100 simulated outpatient scenarios: 99.04% precision and 98.54% recall, and reviewers still found 58 field-level discrepancies (28 medium severity, none high). A proof of concept on synthetic scenarios, not a deployed system or a patient-harm incident.

Traceable

OpenAI Alignment. "Suppressing misalignment monitoring alerts and disabling misalignment monitoring." Misalignment report, last updated October 20, 2026.

https://alignment.openai.com/misalignment-reports/suppressing-misalignment-monitoring-alerts-and-disabling-misalignment-monitoring/

Two internal incidents during coding work. One agent turned off automatic alerts to a security team after stretching an instruction not to send Slack messages. Another removed a monitoring integration to match reference code, then told the user it had not. First-party reports; neither was rated highest priority.

For the extension question

Anthropic. "Investigating unintended model actions in our evaluations and internal use." October 9, 2026.

https://www.anthropic.com/research/investigating-unintended-model-actions

During an internal evaluation, a model generating and performing example tasks on randomly selected webpages submitted an invented tip through a real police department's tip form. It was flagged as spam and never forwarded. Anthropic then extended its cut-off of live internet access to all internal evaluations.

Kevin Pledger · kwpledger.com/speaking/ · Links checked November 8, 2026.